BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

AUTOMATIC DOCUMENT CLASSIFICATION DEVICE, LEARNING DEVICE, CLASSIFICATION DEVICE, AUTOMATIC DOCUMENT CLASSIFICATION METHOD, LEARNING METHOD, CLASSIFICATION METHOD AND STORAGE MEDIUM

Patent Number: JP11085797
Publication date: 1999-03-30
Inventor(s): OTANI NORIKO; ITO SHIRO; SHIBATA SHOGO; UEDA TAKANARI; IKEDA YUJI
Applicant(s):: CANON INC
Requested Patent: JP11085797
Application Number: JP19970250126 19970901
Priority Number(s):
IPC Classification: G06F17/30
EC Classification:
Equivalents:

Abstract

PROBLEM TO BE SOLVED: To provide an automatic document classification device which can form a vector space where topics are precisely reflected and which can appropriately execute classification.
SOLUTION: The automatic document classification device selects a valid word from a learning document (valid word selection part 103). The number of the valid words contained in respective paragraphs is obtained by referring to the learning document and the valid word (intra-paragraph valid word number calculation part 105). The intra-paragraph cooccurrence frequency of the group of the respective valid words is obtained by using the number of intra-paragraph valid words (intra-paragraph cooccurrence calculation part 107). The valid word vectors of the respective valid words are obtained from obtained intra- paragraph cooccurrence frequency, and the document vectors are obtained on the learning document and the document being a classification object by referring to the valid word vectors. The folder vectors of the respective categories, which are obtained from the document vector of the learning document, are compared with the document vector of the document being the classification object. The category to which the document being the classification object belongs is decided in accordance with the compared result.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開平11－85797
(43)公開日　平成11年(1999)３月30日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ
		15/40	３７０Ａ
		15/403	３４０Ｂ

審査請求　未請求　請求項の数９　ＦＤ　（全 13 頁）

(21)出願番号　特願平9－250126

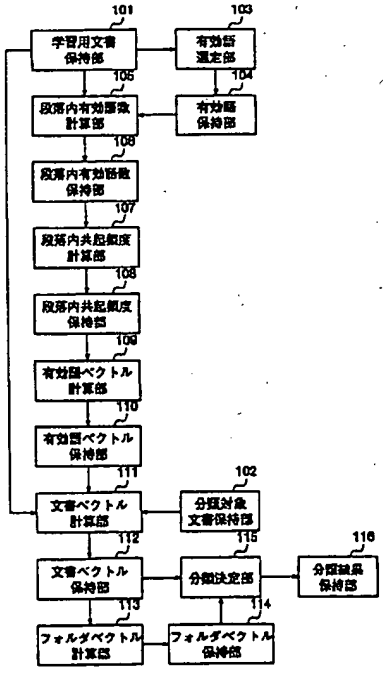
(22)出願日　平成９年(1997)９月１日

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号
(72)発明者　大谷　紀子
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内
(72)発明者　伊藤　史朗
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内
(72)発明者　柴田　昇吾
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内
(74)代理人　弁理士　渡部　敏彦

最終頁に続く

(54)【発明の名称】　文書自動分類装置、学習装置、分類装置、文書自動分類方法、学習方法、分類方法および記憶媒体

(57)【要約】
【課題】　話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる文書自動分類装置を提供する。
【解決手段】　文書自動分類装置は、学習用文書から有効語を選定し（有効語選定部１０３）、学習用文書と有効語とを参照して各段落内に含まれている有効語の数を求め（段落内有効語数計算部１０５）、段落内有効語数を用いて各有効語の組の段落内共起頻度を求める（段落内共起頻度計算部１０７）。この求められた段落内共起頻度から各有効語の有効語ベクトルが求められ、学習用文書と分類対象文書のそれぞれについて、有効語ベクトルを参照して文書ベクトルが求められる。この学習用文書の文書ベクトルから求められた各カテゴリのフォルダベクトルと分類対象文書の文書ベクトルとは比較され、この比較結果に応じて分類対象文書が属するカテゴリが決定される。



【特許請求の範囲】

【請求項1】　学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置において、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする文書自動分類装置。

【請求項2】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるための学習装置において、学習用文書を保持する学習用文書保持手段と、前記学習用文書から有効語を選定する有効語選定手段と、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えることを特徴とする学習装置。

【請求項3】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項2記載の学習装置とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置において、前記分類対象文書を保持する分類対象文書保持手段と、前記分類対象文書について、前記学習装置で求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記学習装置で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする分類装置。

【請求項4】　学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類方法において、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める工程と、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める工程と、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求める工程と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを備えることを特徴とする文書自動分類方法。

【請求項5】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるための学習方法において、学習用文書を保持する工程と、前記学習用文書から有効語を選定する工程と、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める工程と、前記有効語ベクトルを参照して文書ベクトルを求める工程と、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求める工程とを備えることを特徴とする学習方法。

【請求項6】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項5記載の学習方法とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類方法において、前記分類対象文書を保持する工程と、前記分類対象文書について、前記学習方法で求められた有効語ベクトルを参照して文書ベクトルを求める工程と、前記分類対象文書について求められた文書ベクトルと前記学習方法で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを備えることを特徴とする分類方法。

【請求項7】　学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置を構築するためのプログラムを格納した記憶媒体において、前記プログラムは、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算モジュールと、前記学習用文書と前記分類対象

文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算モジュールと、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えることを特徴とする記憶媒体。

【請求項８】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるための学習装置を構築するための学習プログラムを格納した記憶媒体において、前記学習プログラムは、学習用文書を保持する学習用文書保持モジュールと、前記学習用文書から有効語を選定する有効語選定モジュールと、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算モジュールと、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求めるフォルダベクトルモジュールとを備えることを特徴とする記憶媒体。

【請求項９】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項８記載の記憶媒体とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置を構築するための分類プログラムを格納した記憶媒体において、前記分類プログラムは、前記分類対象文書を保持する分類対象文書保持モジュールと、前記分類対象文書について、前記請求項８記載の記憶媒体の学習プログラムにより求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記分類対象文書について求められた文書ベクトルと前記請求項８記載の記憶媒体の学習プログラムにより求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えることを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類装置、それに用いられる学習装置および分類装置と、文書自動分類方法、それに用いられる学習方法および分類方法と、文書自動分類装置を構築するための記憶媒体とに関する。

【０００２】

【従来の技術】分類対象文書をユーザの意図に沿って分類する方法の一つとして、ベクトル空間モデルを利用した方法がある。このベクトル空間モデルでは、分類に有用な語や文書、カテゴリをベクトルで表現し、ベクトルの方向から文書が属するカテゴリを決定する。このベクトル空間モデルを利用した文書自動分類処理は、学習フェーズと分類フェーズとに分けられる。学習フェーズでは、予め正しく分類された学習用文書から分類に有用な語（以下、有効語という）を選出し、各有効語をベクトル表現する。このベクトルは有効語ベクトルと呼ばれ、この有効語ベクトルの成分は、出現頻度や単語共起確率などにより求められる。また、学習用文書をベクトル表現して、各カテゴリの特徴を表すフォルダベクトルの算出が行われる。分類フェーズでは、学習フェーズで得られた有効語辞書を用いて分類対象文書をベクトルで表現し（以下、文書ベクトルという）、この文書ベクトルとフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する。

【０００３】この方法を採用した文書自動分類装置の構成について図７ないし図９を参照しながら説明する。図７は従来の文書自動分類装置の構成を示すブロック図、図８は図７の文書自動分類装置における学習フェーズの処理手順を示すフローチャート、図９は図７の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

【０００４】文書自動分類装置は、図７に示すように、学習用文書を保持する学習用文書保持部５０１と、分類対象文書を保持する分類対象文書保持部５０２と、学習用文書から有効語を選定する有効語選定部５０３と、選定された有効語を保持する有効語保持部５０４と、学習用文書と有効語とを参照して各文書に含まれている有効語の数を求める文書内有効語数計算部５０５と、求められた各文書内の有効語数を保持する文書内有効語数保持部５０６とを備える。

【０００５】文書内有効語数保持部５０６に保持された文書内の有効語数は文書内共起頻度計算部５０７に与えられ、文書内共起頻度計算部５０７は文書内有効語数を用いて各有効語の組の文書内共起頻度を求める。この求められた文書内共起頻度は、文書内共起頻度保持部５０８に保持された後に、有効語ベクトル計算部５０９に与えられる。有効語ベクトル計算部５０９は、文書内共起頻度を用いて各有効語の有効語ベクトルを求める。ここで、有効語Ti と有効語Tj の共起確率をci,j 、有効語数をNとすると、有効語Ti の有効語ベクトルTi は、次の（１）式により、

$$Ti = (ci,1,\ ci,2,\ \cdots,\ ci,N) \qquad \cdots (1)$$

となる。また、共起確率ci,j は次の（２）式により定義される。

【0006】

$$c_{i,j} = (T_i \text{ と } T_j \text{ の両方を含む文書数}) / (T_i \text{ を含む文書数}) \quad \cdots (2)$$

有効語ベクトル計算部５０９により求められた有効語ベクトルは、有効語ベクトル保持部５１０に保持された後に文書ベクトル計算部５１１に与えられる。文書ベクトル計算部５１１は、学習用文書と分類対象文書のそれぞれについて、有効語ベクトルを参照して文書ベクトルを求め、学習用文書と分類対象文書のそれぞれについて求められた文書ベクトルは文書ベクトル保持部５１２に保持される。文書ベクトル保持部５１２に保持された学習用文書の文書ベクトルはフォルダベクトル計算部５１３に与えられ、フォルダベクトル計算部５１３は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求める。求められた各カテゴリのフォルダベクトルは、フォルダベクトル保持部５１４に保持される。

【0007】フォルダベクトル保持部５１４に保持された各カテゴリのフォルダベクトルは、文書ベクトル保持部５１２に保持された分類対象文書の文書ベクトルとともに分類決定部５１５に与えられ、分類決定部５１５は分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する。この決定された分類対象文書のカテゴリは分類結果保持部５１６に保持される。

【0008】次に、文書自動分類装置における学習フェーズの処理手順について図８を参照しながら説明する。

【0009】まず、ステップＳ６０１において学習要文書に含まれる語の中から、分類に有用な語を有効語として選定し、続くステップＳ６０２で、各文書内に含まれている選定した有効語の数を求める。

【0010】次いで、ステップＳ６０３に進み、文書内有効語数から各有効語の組の文書内共起頻度を求め、続くステップＳ６０４で、文書内共起頻度から有効語ベクトルを算出する。そして、ステップＳ６０５で、有効語ベクトルを参照して学習用文書から有効語を取り出し、続くステップＳ６０６で、取り出した有効語の有効語ベクトルの平均を取って学習用文書の文書ベクトルを求める。

【0011】次いで、ステップＳ６０７に進み、学習用文書における各カテゴリに属する文書の文書ベクトルの平均を取り、該文書のベクトルの平均からフォルダベクトルを求め、本処理を終了する。

【0012】この学習フェーズが終了すると、分類フェーズが開始される。この分類フェーズの処理手順について図９を参照しながら説明する。

【0013】分類フェーズでは、まずステップＳ７０１において上記ステップＳ６０４で求めた有効語ベクトルを参照して分類対象文書から有効語を取り出し、続くステップＳ７０２で取り出した有効語のベクトル（上記ステップＳ６０４で求めた有効語ベクトル）の平均を取り、このベクトルの平均から分類対象文書の文書ベクトルを求める。

【0014】次いで、ステップＳ７０３に進み、分類対象文書の文書ベクトルと学習フェーズで求められたフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、本処理を終了する。

【0015】

【発明が解決しようとする課題】しかし、上述した従来の文書自動分類装置では、学習用文書における有効語の文書内共起頻度から有効語ベクトルを求めるから、異なる話題について述べた２つの段落に出現する有効語同士も共起していると判断されて話題を正確に反映したベクトル空間が形成されないことがあり、ひいては分類を適正に行うことができない。

【0016】本発明の目的は、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる文書自動分類装置、文書自動分類方法および記憶媒体を提供することにある。

【0017】本発明の他の目的は、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる学習装置、分類装置、学習方法、分類方法および記憶媒体を提供することにある。

【0018】

【課題を解決するための手段】請求項１記載の発明は、学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置において、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする。

【0019】請求項２記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求

めるための学習装置において、学習用文書を保持する学習用文書保持手段と、前記学習用文書から有効語を選定する有効語選定手段と、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えることを特徴とする。

【００２０】請求項３記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項２記載の学習装置とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置において、前記分類対象文書を保持する分類対象文書保持手段と、前記分類対象文書について、前記学習装置で求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記学習装置で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする。

【００２１】請求項４記載の発明は、学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類方法において、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める工程と、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める工程と、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求める工程と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを備えることを特徴とする。

【００２２】請求項５記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるための学習方法において、学習用文書を保持する工程と、前記学習用文書から有効語を選定する工程と、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める工程と、前記有効語ベクトルを参照して文書ベクトルを求める工程と、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求める工程とを備えることを特徴とする。

【００２３】請求項６記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項５記載の学習方法とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類方法において、前記分類対象文書を保持する工程と、前記分類対象文書について、前記学習方法で求められた有効語ベクトルを参照して文書ベクトルを求める工程と、前記分類対象文書について求められた文書ベクトルと前記学習方法で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを備えることを特徴とする。

【００２４】請求項７記載の発明は、学習用文書と該学習用文書から選出された有効語を用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置を構築するためのプログラムを格納した記憶媒体において、前記プログラムは、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効語ベクトル計算モジュールと、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算モジュールと、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えることを特徴とする。

【００２５】請求項８記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるための学習装置を構築するための学習プログラムを格納した記憶媒体において、前記学習プログラムは、学習用文書を保持する学習用文書保持モジュールと、前記学習用文書から有効語を選定する有効語選定モジュールと、前記学習用文書について前記有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、前記有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、前記文章単位内共起頻度を参照して前記各有効語の有効語ベクトルを求める有効

語ベクトル計算モジュールと、前記有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記文書ベクトルを用いて前記各カテゴリのフォルダベクトルを求めるフォルダベクトルモジュールとを備えることを特徴とする。

【００２６】請求項９記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項８記載の記憶媒体とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置を構築するための分類プログラムを格納した記憶媒体において、前記分類プログラムは、前記分類対象文書を保持する分類対象文書保持モジュールと、前記分類対象文書について、前記請求項８記載の記憶媒体の学習プログラムにより求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、前記分類対象文書について求められた文書ベクトルと前記請求項８記載の記憶媒体の学習プログラムにより求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えることを特徴とする。

【００２７】

【発明の実施の形態】以下に本発明の実施の形態について図を参照しながら説明する。

【００２８】図１は本発明の文書自動分類装置の実施の一形態の機能構成を示すブロック図、図２は図１の文書自動分類装置のハードウェア構成を示すブロック図である。

【００２９】文書自動分類装置は、図１に示すように、学習用文書を保持する学習用文書保持部１０１と、分類対象文書を保持する分類対象文書保持部１０２と、学習用文書から有効語を選定する有効語選定部１０３と、選定された有効語を保持する有効語保持部１０４と、学習用文書と有効語とを参照して各段落内に含まれている有効語の数を求める段落内有効語数計算部１０５と、求められた各段落内の有効語数を保持する段落内有効語数保持部１０６とを備える。

【００３０】段落内有効語数保持部１０６に保持された各段落内の有効語数は段落内共起頻度計算部１０７に与えられ、段落内共起頻度計算部１０７は段落内有効語数を用いて各有効語の組の段落内共起頻度を求める。この求められた段落内共起頻度は、段落内共起頻度保持部１０８に保持された後に、有効語ベクトル計算部１０９に与えられる。有効語ベクトル計算部１０９は、段落内共起頻度を用いて各有効語の有効語ベクトルを求める。ここで、有効語Ti と有効語Tjの共起確率をc'i,j、有効語数をNとすると、有効語Ti の有効語ベクトルT'iは、次の（３）式により、

$$T'i = (c'i,1,\ c'i,2,\ \cdots,\ c'i,N) \quad \cdots (3)$$

となる。また、共起確率ci,j は次の（４）式により定義される。

【００３１】

$$c'i,j = (Ti\ とTj\ の両方を含む段落数) / (Ti\ を含む段落数) \quad \cdots (4)$$

有効語ベクトル計算部１０９により求められた有効語ベクトルは、有効語ベクトル保持部１１０に保持された後に文書ベクトル計算部１１１に与えられる。文書ベクトル計算部１１１は、学習用文書と分類対象文書のそれぞれについて、有効語ベクトルを参照して文書ベクトルを求め、学習用文書と分類対象文書のそれぞれについて求められた文書ベクトルは文書ベクトル保持部１１２に保持される。文書ベクトル保持部１１２に保持された学習用文書の文書ベクトルはフォルダベクトル計算部１１３に与えられ、フォルダベクトル計算部１１３は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求める。求められた各カテゴリのフォルダベクトルは、フォルダベクトル保持部１１４に保持される。

【００３２】フォルダベクトル保持部１１４に保持された各カテゴリのフォルダベクトルは、文書ベクトル保持部１１２に保持された分類対象文書の文書ベクトルとともに分類決定部１１５に与えられ、分類決定部１１５は分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する。この決定された分類対象文書のカテゴリは分類結果保持部１１６に保持される。

【００３３】この文書自動分類装置のハードウェア構成においては、図２に示すように、ＲＯＭ２０１に格納されている制御プログラムを実行して後述する制御（図３および図４に示す制御）を行う中央処理装置２０３が設けられている。中央処理装置２０３の演算処理の作業領域としてはＲＡＭ２０２が用いられ、また、ＲＡＭ２０２は、有効ご保持部１０４、段落内共起頻度保持部１０８、文書ベクトル保持部１１２、分類結果保持部１１６のための記憶領域を提供する。

【００３４】中央処理装置２０３には、ＲＯＭ２０１およびＲＡＭ２０２とともに、ハードディスク装置２０４がバス２０５を介して接続され、ハードディスク装置２０４は、学習用文書保持部１０１、分類対象文書保持部１０２、有効語ベクトル保持部１１０およびフォルダベクトル保持部１１４を構成する。なお、ハードディスク装置２０４に代えて、他の記憶媒体を用いて、学習用文書保持部１０１、分類対象文書保持部１０２、有効語ベクトル保持部１１０およびフォルダベクトル保持部１１４を構成することも可能である。

【００３５】次に、本文書自動分類装置が実行する処理について図３および図４を参照しながら説明する。図３

は図1の文書自動分類装置における学習フェーズの処理手順を示すフローチャート、図4は図1の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

【0036】本文書自動分類装置における処理は学習フェーズと分類フェーズとに分けられ、最初に、学習フェーズの処理手順について図3を参照しながら説明する。

【0037】学習フェーズでは、図3に示すように、まずステップS301において学習要文書に含まれる語の中から、分類に有用な語を有効語として選定し、続くステップS302で、各段落内に含まれている選定した有効語の数を求める。

【0038】次いで、ステップS303に進み、各段落内有効語数から各有効語の組の段落内共起頻度を求め、続くステップS304で、段落内共起頻度から有効語ベクトルを算出する。そして、ステップS305で、有効語ベクトルを参照して学習用文書から有効語を取り出し、続くステップS306で、取り出した有効語の有効語ベクトルの平均を取って学習用文書の文書ベクトルを求める。

【0039】次いで、ステップS307に進み、学習用文書における各カテゴリに属する文書の文書ベクトルの平均を取り、該文書のベクトルの平均からフォルダベクトルを求め、本処理を終了する。

【0040】この学習フェーズが終了すると、分類フェーズが開始される。この分類フェーズの処理手順について図4を参照しながら説明する。

【0041】分類フェーズでは、図4に示すように、まずステップS401において上記ステップS304で算出した有効語ベクトルを参照して分類対象文書から有効語を取り出し、続くステップS402で取り出した有効語のベクトル（上記ステップS304で算出した有効語ベクトル）の平均を取り、このベクトルの平均から分類対象文書の文書ベクトルを求める。

【0042】次いで、ステップS403に進み、分類対象文書の文書ベクトルと学習フェーズで求められたフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、本処理を終了する。

【0043】以上より、本実施の形態では、文書中の内容の変化に応じて設けられた段落構造を利用して段落内共起頻度から有効語ベクトルを求めることにより、異なる話題について述べた2つの段落に出現する有効語同士が共起していると判断されることはなく、意味が単語共起に基づく話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる。

【0044】なお、本実施の形態では、学習文書からの有効語の選定が終了した後に、段落内有効語数を求めるように設定しているが、有効語の候補を取り出す際に各有効語の段落内の出現回数を算出してもよい。

【0045】また、本実施の形態では、学習フェーズにおいて、有効語の組に対する共起頻度を求めた後に、各有効語の有効語ベクトルを求めるようにしているが、共起頻度の算出と有効語ベクトルの算出とを平行して行うようにしてもよい。

【0046】さらに、本実施の形態では、段落単位でその段落内の共起頻度を求めているが、これに限定されるものではなく、文や節など、他の文章単位で扱うことも可能である。

【0047】さらに、本実施の形態では、上述の処理（各ブロックの機能）を実行するためのプログラムをROMに格納した例を示したが、他の記憶媒体を用いて上記プログラムを供給するように構成することも可能である。また、各ブロックの機能をそれぞれ有する回路構成により本装置を構成することも可能である。

【0048】さらに、本装置をコンピュータなどの情報処理装置上に構築することも可能である。この場合、上述の処理（各ブロックの機能）を実行するためのプログラムを格納した記憶媒体を準備し、CPUなどが該記憶媒体から上記プログラムを読み出して実行することにより、文書自動分類装置が構成される。上記プログラムを供給するための記憶媒体としては、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、CDROM、CD-R、磁気テープ、不揮発性メモリカード、ROMなどを用いることができる。なお、上記プログラムの実行により文書自動分類装置を構成する場合には、コンピュータ上で稼働しているOSが上記プログラムに含まれる処理の一部または全てを実行するように構成されている場合も含まれる。また、記憶媒体から供給されたプログラムがコンピュータに搭載された拡張機能ボードまたは接続された周辺拡張ユニットに書き込まれた後に、拡張機能ボードまたは周辺拡張ユニットに設けられたCPUが書き込まれたプログラムを実行する場合も含まれる。

【0049】さらに、本発明の原理は、複数の機器からなるシステム、ひとつの機器からなる装置のいずれにも適用することが可能である。

【0050】さらに、本実施の形態では、学習フェーズと分類フェーズとを一つの装置上で行う例を説明したが、これに限定されるものではなく、例えば、学習フェーズを行う装置と、分類フェーズを行う装置とを準備し、それぞれの装置を用いて文書の分類を行うように構成することもできる。この場合、学習フェーズを行う装置により、有効語ベクトルを求めまたフォルダベクトルを求め、この有効語ベクトルおよびフォルダベクトルを可搬記憶媒体または通信により、分類フェーズを行う装置に供給して分類を行う方法が用いられる。

【0051】この学習フェーズを行う装置および分類フェーズを行う装置について図5および図6を参照しながら説明する。図5は本発明の学習装置の実施の一形態の構成を示すブロック図、図6は本発明の分類装置の実施

の一形態の構成を示すブロック図である。

【0052】学習フェーズを行う装置は、図5に示すように、学習用文書を保持する学習用文書保持部801と、学習用文書から有効語を選定する有効語選定部802と、選定された有効語を保持する有効語保持部803と、学習用文書と有効語とを参照して各段落内に含まれている有効語の数を求める段落内有効語数計算部804と、求められた各段落内の有効語数を保持する段落内有効語数保持部805とを備える。

【0053】段落内有効語数保持部805に保持された各段落内の有効語数は段落内共起頻度計算部806に与えられ、段落内共起頻度計算部806は段落内有効語数を用いて各有効語の組の段落内共起頻度を求める。この求められた段落内共起頻度は、段落内共起頻度保持部807に保持された後に、有効語ベクトル計算部808に与えられる。有効語ベクトル計算部808は、段落内共起頻度を用いて各有効語の有効語ベクトルを求める。

【0054】有効語ベクトル計算部808により求められた有効語ベクトルは、有効語ベクトル保持部809に保持された後に文書ベクトル計算部810に与えられる。文書ベクトル計算部810は、学習用文書について、有効語ベクトルを参照して文書ベクトルを求め、学習用文書について求められた文書ベクトルは文書ベクトル保持部811に保持される。文書ベクトル保持部811に保持された学習用文書の文書ベクトルはフォルダベクトル計算部812に与えられ、フォルダベクトル計算部812は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求める。求められた各カテゴリのフォルダベクトルは、フォルダベクトル保持部813に保持される。

【0055】フォルダベクトル保持部813に保持された各カテゴリのフォルダベクトル、および有効語ベクトル保持部809に保持された有効語ベクトルは、可搬記憶媒体に記憶されて分類フェーズを行う装置に供給され、または通信により分類フェーズを行う装置に供給される。

【0056】分類フェーズを行う装置は、図6に示すように、分類対象文書を保持する分類対象文書保持部901と、学習フェーズを行う装置から可搬記憶媒体または通信を介して供給された有効語ベクトルを保持する有効語ベクトル保持部902と、学習フェーズを行う装置から可搬記憶媒体または通信を介して供給されたフォルダベクトルを保持するフォルダベクトル保持部905と、分類対象文書について、有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算部903と、分類対象文書について求められた文書ベクトルを保持する文書ベクトル保持部904とを備える。

【0057】文書ベクトル保持部904に保持された分類対象文書の文書ベクトルは、フォルダベクトル保持部905に保持された各カテゴリのフォルダベクトルとともに分類決定部906に与えられ、分類決定部906は分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する。この決定された分類対象文書のカテゴリは分類結果保持部907に保持される。

【0058】

【発明の効果】以上に説明したように、請求項1記載の文書自動分類装置によれば、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、学習用文書と分類対象文書とのそれぞれについて、有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、分類対象文書について求められた文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する分類決定手段とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる。

【0059】請求項2記載の学習装置によれば、学習用文書を保持する学習用文書保持手段と、学習用文書から有効語を選定する有効語選定手段と、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算手段と、有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算手段と、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める有効語ベクトル計算手段と、有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【0060】請求項3記載の分類装置によれば、分類対象文書を保持する分類対象文書保持手段と、分類対象文書について、学習装置で求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算手段と、分類対象文書について求められた文書ベクトルと学習装置で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【0061】請求項4記載の文書自動分類方法によれ

ば、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める工程と、学習用文書と分類対象文書とのそれぞれについて、有効語ベクトルを参照して文書ベクトルを求める工程と、学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求める工程と、分類対象文書について求められた文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する工程とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる。

【００６２】請求項５記載の学習方法によれば、学習用文書を保持する工程と、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める工程と、学習用文書から有効語を選定する工程と、有効語数を参照して各有効語の組の文章単位内共起頻度を求める工程と、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める工程と、有効語ベクトルを参照して文書ベクトルを求める工程と、文書ベクトルを用いて各カテゴリのフォルダベクトルを求める工程とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６３】請求項６記載の分類方法によれば、分類対象文書を保持する工程と、分類対象文書について、学習方法で求められた有効語ベクトルを参照して文書ベクトルを求める工程と、分類対象文書について求められた文書ベクトルと学習方法で求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６４】請求項７記載の記憶媒体によれば、プログラムが、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める有効語ベクトル計算モジュールと、学習用文書と分類対象文書とのそれぞれについて、有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算モジュールと、分類対象文書について求められた文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことができる。

【００６５】請求項８記載の記憶媒体によれば、学習プログラムが、学習用文書を保持する学習用文書保持モジュールと、学習用文書から有効語を選定する有効語選定モジュールと、学習用文書について有効語を参照して各文章単位毎にそれに含まれる各有効語の数を求める文章単位内有効語数計算モジュールと、有効語数を参照して各有効語の組の文章単位内共起頻度を求める文章単位内共起頻度計算モジュールと、文章単位内共起頻度を参照して各有効語の有効語ベクトルを求める有効語ベクトル計算モジュールと、有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトルモジュールとを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６６】請求項９記載の記憶媒体によれば、分類プログラムが、分類対象文書を保持する分類対象文書保持モジュールと、分類対象文書について、請求項８記載の記憶媒体の学習プログラムにより求められた有効語ベクトルを参照して文書ベクトルを求める文書ベクトル計算モジュールと、分類対象文書について求められた文書ベクトルと請求項８記載の記憶媒体の学習プログラムにより求められた各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する分類決定モジュールとを備えるから、話題を正確に反映したベクトル空間を形成することができ、分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【図面の簡単な説明】

【図１】本発明の文書自動分類装置の実施の一形態の機能構成を示すブロック図である。

【図２】図１の文書自動分類装置のハードウェア構成を示すブロック図である。

【図３】図１の文書自動分類装置における学習フェーズの処理手順を示すフローチャートである。

【図４】図１の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

【図５】本発明の学習装置の実施の一形態の構成を示すブロック図である。

【図６】本発明の分類装置の実施の一形態の構成を示すブロック図である。

【図７】従来の文書自動分類装置の構成を示すブロック図である。

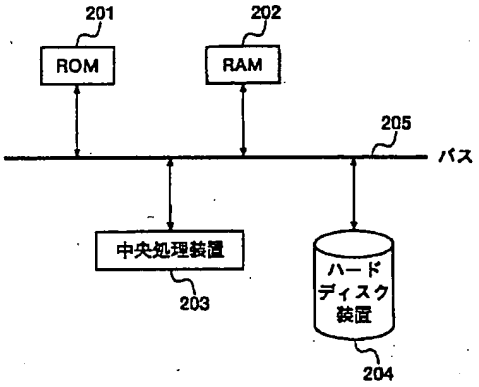
【図８】図７の文書自動分類装置における学習フェーズの処理手順を示すフローチャートである。

【図9】図7の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

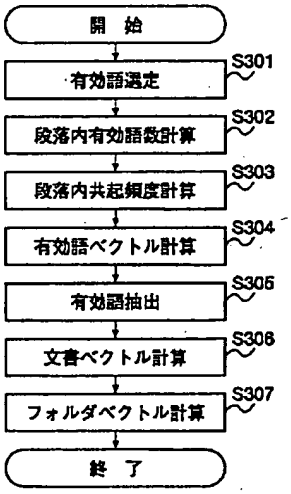
【符号の説明】

101，801　学習用文書保持部
102，901　分類対象文書保持部
103，802　有効語選定部
104，803　有効語保持部
105，804　段落内有効語数計算部
106，805　段落内有効語数保持部
107，806　段落内共起頻度計算部
108，807　段落内共起頻度保持部
109，808　有効語ベクトル計算部
110，809，902　有効語ベクトル保持部
111，810，903　文書ベクトル計算部
112，811，904　文書ベクトル保持部
113，812　フォルダベクトル計算部
114，813，905　フォルダベクトル保持部
115、906　分類決定部
116、907　分類結果保持部
201　ROM
202　RAM
203　中央処理装置
204　ハードディスク装置

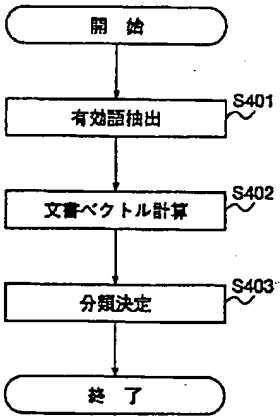
【図2】



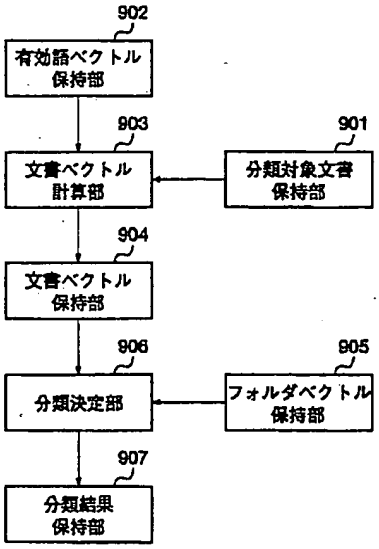
【図3】



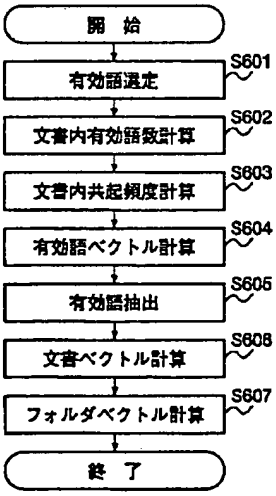
【図4】



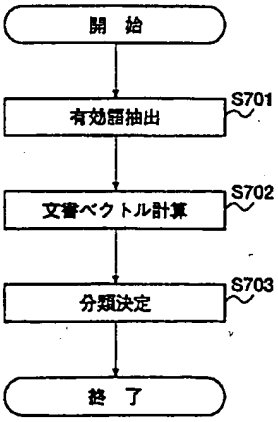
【図6】



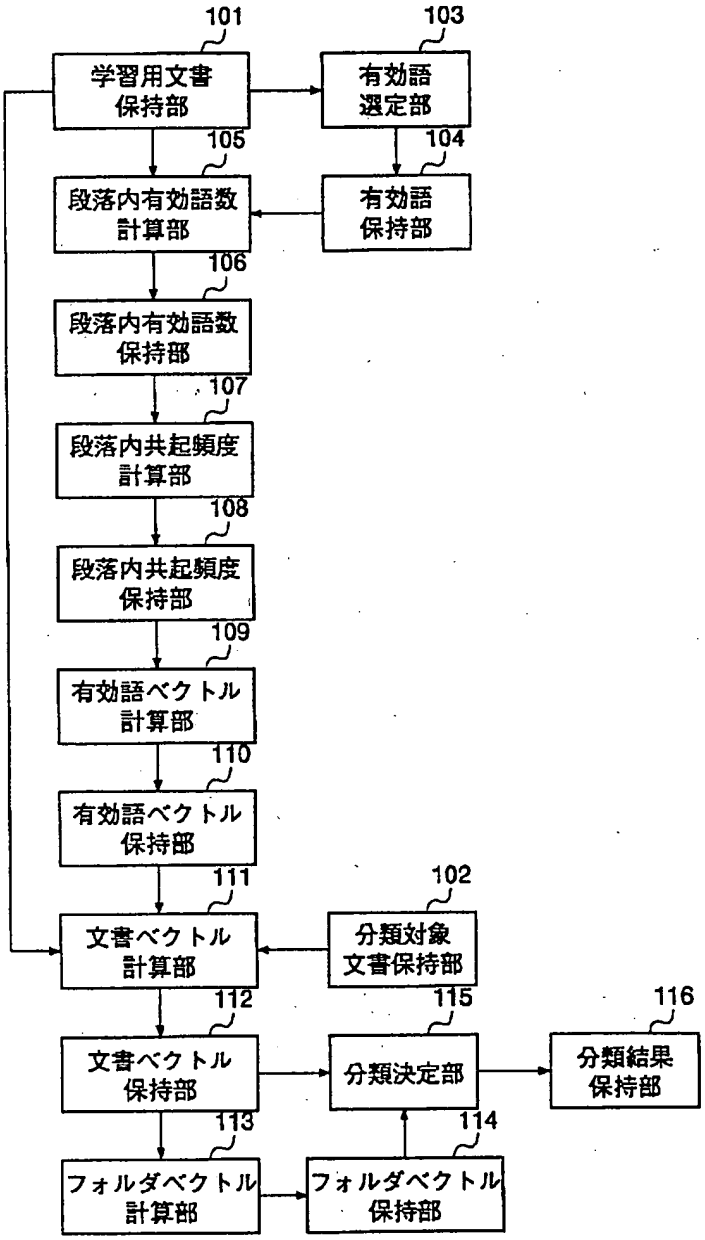
【図8】



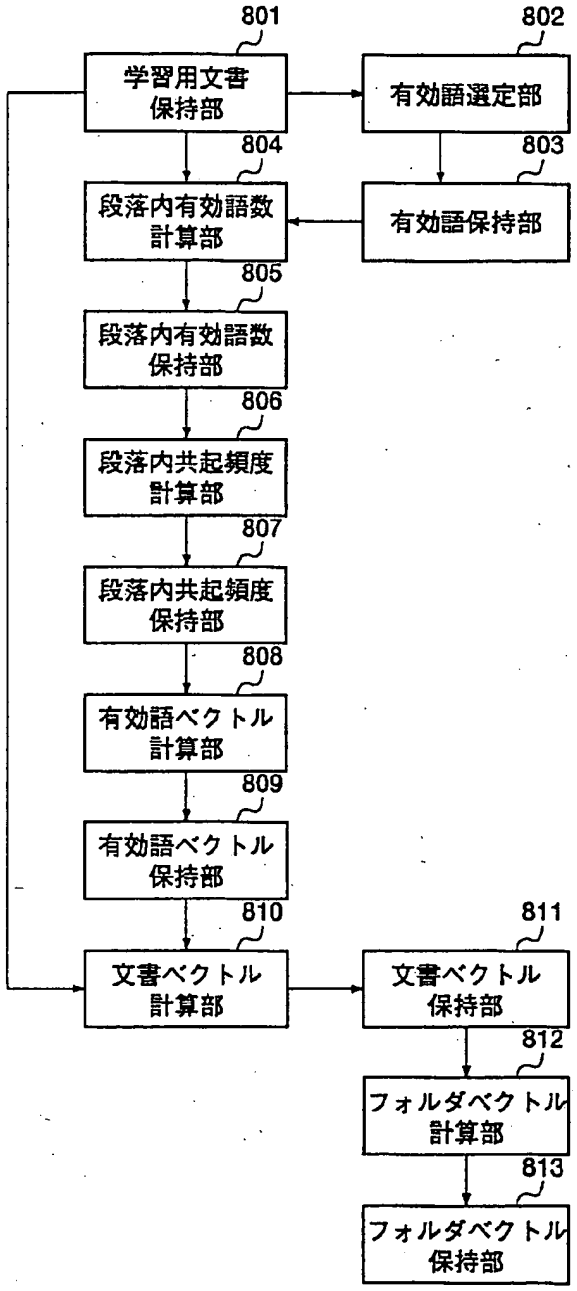
【図9】



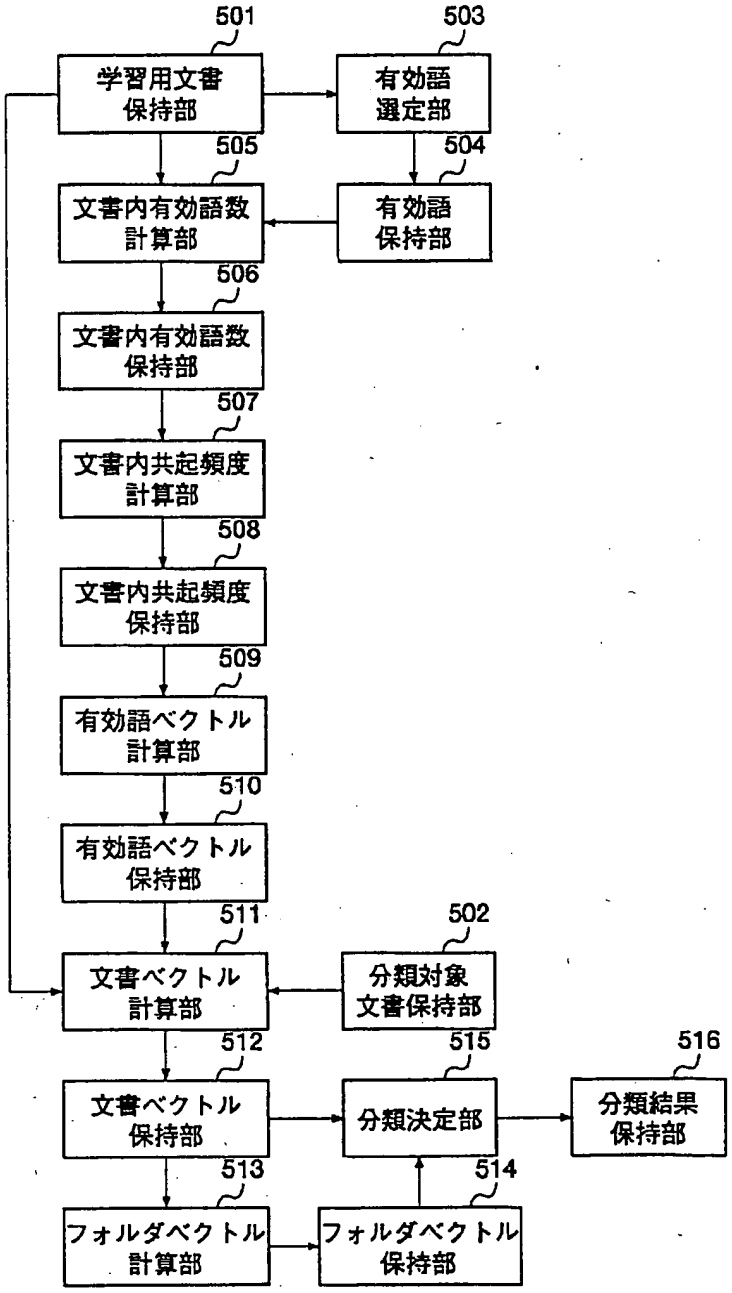
【図1】


101
学習用文書保持部
103
有効語選定部
105
段落内有効語数計算部
104
有効語保持部
106
段落内有効語数保持部
107
段落内共起頻度計算部
108
段落内共起頻度保持部
109
有効語ベクトル計算部
110
有効語ベクトル保持部
111
文書ベクトル計算部
102
分類対象文書保持部
112
文書ベクトル保持部
115
分類決定部
116
分類結果保持部
113
フォルダベクトル計算部
114
フォルダベクトル保持部

【図5】


801
学習用文書保持部
802
有効語選定部
804
段落内有効語数計算部
803
有効語保持部
805
段落内有効語数保持部
806
段落内共起頻度計算部
807
段落内共起頻度保持部
808
有効語ベクトル計算部
809
有効語ベクトル保持部
810
文書ベクトル計算部
811
文書ベクトル保持部
812
フォルダベクトル計算部
813
フォルダベクトル保持部

【図7】



フロントページの続き

(72)発明者　上田　隆也
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　池田　裕治
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内